(54) Title: WORD SENSE DISAMBIGUATION

	Doc #1	Doc #2	. . .	Doc #n
able	1	0	. . .	1
acid	0	2	. . .	0
. . .	. . .	. . .	. . .	. . .
<term m>	1	0	. . .	1

Term by Document Matrix

(57) Abstract: Systems and methods for word sense disambiguation, including discerning one or more senses or occurrences, distinguishing between senses or occurrences, and determining a meaning for a sense or occurrence of a subject term. In a collection of documents containing terms and a reference collection containing at least one meaning associated with a term, the method includes forming a vector space representation of terms (12) and documents (14) as elements (16) within a matrix (10). In some embodiments, the vector space is a latent semantic index vector space. In some embodiments, occurrences are clustered to discern or distinguish a sense of a term. In preferred embodiments, meaning of a sense or occurrence is assigned based on either correlation with an external reference source, or proximity to a reference source that has been indexed into the space.

WORD SENSE DISAMBIGUATION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims benefit of, and hereby incorporates in its entirety, U.S.
5   Provisional Application No. 60/227,301, filed August 24, 2000.

FIELD OF THE INVENTION

The present invention relates to the field of processing machine-readable documents, and in particular to a method and system for word sense disambiguation.

10

BACKGROUND OF THE DISCLOSURE

One of the most challenging problems in information technology today is improving the methods through which people interact with computer systems. This problem is made difficult because, among other reasons, humans typically rely
15   overwhelmingly on natural language, e.g., English, for transfer of information. Although people deal with natural language with relative ease, processing of natural language using computer systems has turned out to be an exceedingly difficult problem. One source of this difficulty is the ambiguous meaning of many natural language terms.

There are several sources of ambiguity in language; however, the most pervasive
20   is ambiguity in term meaning. This is referred to as *polysemy*. In general, all human languages exhibit a high degree of polysemy.

In English, most frequently used terms have several common meanings. For example, the word *fire* can mean: a combustion activity; to terminate employment; to launch; or to excite (as in fire up). For the 200 most-polysemous terms in English, the
25   typical verb has more than 12 common meanings, or senses. The typical noun from this set has more than eight common senses. For the 2000 most-polysemous terms in English, the typical verb has more than eight common senses and the typical noun has more than five.

Polysemy presents a major obstacle for all computer systems that attempt to deal with human language. This is true for both written and spoken language. In order to achieve commercially acceptable performance, some means must be found to reliably discern the presence of a word sense, reliably distinguish between different senses of the same term, and reliably determine the correct sense for the terms that are encountered. Such discerning, distinguishing, or determining of the sense of terms in written or spoken language is referred to, generally, as *word sense disambiguation* (WSD).

The effects of polysemy in computer processing of language are ubiquitous. The March 1998 issue of the journal *Computational Linguistics* is a special issue on word sense disambiguation. In their introduction to the issue, Ide and Veronis state: "Sense disambiguation ... is necessary at one level or another to accomplish most natural language processing tasks. It is obviously essential for language understanding applications, such as man-machine communication; it is at least helpful, and in some instances required, for applications whose aim is not language understanding." Ide, N., and Veronis, J., Introduction to the Special Issue on Word Sense Disambiguation: the State of the Art, Special Issue on Word Sense Disambiguation, *Computational Linguistics*, Volume 24, #1, March, 1998, pp 1-40.

The importance of word sense disambiguation can be seen in the case of machine translation systems. An incorrect choice of sense for even a single term can change the meaning of a passage. Such misinterpretation can yield incorrect results. A well-known example is the (perhaps apocryphal) instance of a machine translation system that reportedly mistranslated *hydraulic rams* as *water goats*. Although there are other sources of ambiguity, polysemy presents a major problem in machine translation today.

Polysemy also is a critical problem in the area of information retrieval. For example, a query to a database on the word *strike* in order to find information on labor disputes would also be inundated with information on the following:

- sports (strikes in bowling and baseball);

- military actions (air strikes, etc.);
- objects that strike one another;
- striking of matches and oil;
- cases where people strike up a conversation;
5
- events that strike people as funny; and
- a range of other topics in which this particularly polysemous term occurs.

Polysemy creates difficulties in all aspects of text filtering and retrieval, including query interpretation, categorization, summarization, result matching, and ranking.

Polysemy limits the performance of systems in other application areas, including
10   but not limited to:

- speech recognition;
- text-to-speech conversion;
- content and thematic analysis;
- grammatical analysis;
15
- text categorization;
- text summarization; and
- natural language understanding.

In each of these applications, differentiating among multiple word senses is a key
to improving performance. In some applications, it is sufficient to be able to identify
20   which occurrences of a given term have the same sense. That is, it is sufficient to be able
to associate each occurrence of a given term with the collection of occurrences of that
term that has the same sense. This is referred to as *sense discrimination*. In other
applications, each such collection is labeled with the specific meaning of the term that
corresponds to that cluster. This is referred to as *sense tagging*. The present invention
25   addresses each of these objectives.

The importance of word sense disambiguation was noted in one of the first papers ever published on computer processing of language. Weaver, W., translation, mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, W, and Booth, A. D., (Eds.), Machine Translation of Languages, John Wiley & Sons, New York, 1955, pp. 15-23.

5 Since that time, authorities in the area of natural language understanding and related fields have consistently cited both the importance of the topic and the significant limitations of available techniques. In the published version of his doctoral dissertation (1989), Cottrell noted: "Lexical ambiguity ... is perhaps the most important problem facing an NLU (Natural Language Understanding) system." Cotrell, G.W., A

10 Connectionist Approach to Word Sense Disambiguation, Morgan Kaufman Publishers, 1989. In their 1997 SIGLEX presentation, Resnik and Yarowsky declared: "Word sense disambiguation is perhaps the great open problem at the lexical level of natural language processing." Resnick, P., and Yarowski, D., A Perspective on Word Sense Disambiguation Methods and their Evaluation, in: Proceedings of SIGLEX '97,

15 Washington, D.C., 1997, pp. 79-86. In his 1997 paper, Hwee similarly noted: "WSD is a challenging task and much improvement is still needed." Hwee, T.N., Getting Serious about Word Sense Disambiguation, in: Tagging Text with Lexical Semantics: Why, What, and How?, workshop, April 4-5, 1997, Washington, D.C. In their introduction to the March, 1998 special issue of the journal *Computational Linguistics* devoted to the

20 topic of word sense disambiguation, Ide and Veronis state: "the problem of word sense disambiguation has taken center stage, and it is frequently cited as one of the most important problems in natural language processing research today". In the summary to their introduction, they state: "in the broad sense, relatively little progress seems to have been made in nearly 50 years". In the conclusion to a chapter on word sense

25 disambiguation in their landmark 1999 book on natural language processing, Manning and Schutze note: "Much research remains to be done on word sense disambiguation". Manning, C., and Schutze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999.

Some approaches to word sense disambiguation rely upon the existence of dictionaries or thesauri that contain the terms to be disambiguated. In these methods, the sense of a given term is estimated based upon the terms in a window surrounding the given term. The sense, or meaning, is chosen based upon the correlation of these surrounding terms with the terms that happen to be used in the various dictionary definitions or thesaurus entries. The accuracy of these approaches has been poor, in the 50% to 70% range, even for terms with small numbers of distinctly differentiated senses. Accuracy can be improved in those specific cases where a term to be disambiguated appears more than once in a text segment of interest. This is done through employing an assumption that all such occurrences correspond to the same sense of that term (the "one sense per discourse" assumption.) The applicability of this condition is limited, however, by the fact that term occurrences follow a hyperbolic distribution (Zipf's Law). That is, on average, more than half of all terms that appear in a given passage of text of reasonable length will occur only once in that passage.

These approaches to disambiguation are fundamentally limited by the vocabulary coverage of the dictionaries and thesauri employed. The approaches exhibit significantly degraded performance in domains that employ specialized terminology. The static nature of the dictionaries and thesauri used also make them inherently unsuited for dealing with changes in language over time.

Approaches have been proposed for automatic (statistical) extension of thesauri. Thesaurus-based techniques, however, fundamentally are unable to deal with word senses that are not strongly correlated with the specific categories used in the thesaurus.

Other approaches to word sense disambiguation exploit the fact that multiple senses of a term in a given language may translate as distinct terms in another language. These approaches require a large, carefully translated parallel corpus together with a bilingual dictionary for the language pair in question. Generation of such corpora is resource-intensive and is inherently constrained in terms of domain. For a given term to be disambiguated, the technique requires that the specific phrase in which that term

occurs also be present in the parallel corpora. Thus, either very large corpora are required or many term occurrences will not be amenable to disambiguation. In addition, there is rarely a perfect one-to-one mapping between word sense in one language and distinct terms in another language. Thus, there are many senses that cannot be

5   disambiguated even if the exact phrase is found. Other senses will yield only the most probable sense among many. Often these probabilities will be less than 50%.

Prior statistical approaches to word sense disambiguation typically have required the existence of a training set in which the terms of interest are labeled as to sense. The cost of creating such training sets has severely limited the application of these

10   approaches. Such approaches are very limited in domain and are not well suited for dealing with the changes in language over time. In addition, these approaches invariably employ a fixed context window (typically plus or minus a few terms), which significantly compromises performance. Some approaches to statistical disambiguation have employed unsupervised techniques. The advantage of this approach is that the training

15   set does not need to be labeled as to sense. These approaches are only applicable to sense discrimination. The effectiveness of the technique is highly dependent upon the correlation between the terms and their usage in the training set and the term distribution and usage in the text to be disambiguated. The technique also is weak for infrequently used senses and senses that have few distinctive collocated terms. Performance of these

20   approaches has been reported as 5% to 10% lower than that of dictionary-based approaches.

One class of approaches to word sense disambiguation relies on the parsing of sentences. Accurate parsing for this purpose depends upon the prior existence of a world model. These approaches suffer from two drawbacks: first, they require a high degree of

25. manual effort in constructing the necessary world models; and second, tractable world models are inherently limited in the extent of the domain that they can cover.

Other approaches to WSD rely on analysis of sentences based on sets of processing rules. In order to obtain even modest accuracy these collections of rules

become elaborate, involving templates such as *Agent-Action-Object* triples. An Agent-Action-Object triple is a set of three terms that can stand in the relationship of an *agent* that carries out the specified *action* on the specified *object*. Once again, these approaches require large amounts of initial manual work and are inherently limited in the scope of their applicability. Some such systems attempt to achieve reasonable coverage and accuracy through use of *word experts* — handcrafted rule-based expert systems for each term to be encountered. Information is exchanged among these word experts on an iterative basis until a plausible estimate of sentence meaning is generated. The complexity of such systems is indicative of the lengths to which people have been willing to go in order to obtain useful levels of word sense disambiguation. The individual word experts in such systems can constitute pages of code. Even with such elaboration, however, the rapidly increasing combinatorics significantly limits the extent of the context that can be taken into account.

Connectionist approaches to word sense disambiguation suffer many of the same drawbacks as the rule-based approaches. Such approaches typically require a large amount of manual effort required in establishing the networks. As the extent of the context that is taken into account is increased, the complexity of the networks grows rapidly. Thus, both domain of applicability and extent of context are constrained.

A large-scale evaluation of existing word sense disambiguation systems has been conducted under the auspices of ACL-SIGLEX, EURALEX, ELSNET, and the EU projects ECRAN and SPARKLE. This project entailed testing of 19 different approaches to the problem. University of Brighton (England), Information Technology Research Institute. *SENSEVAL*, Evaluating Word Sense Disambiguation Systems [online]. 2001 [retrieved on 2001-08-10]. Retrieved from the Internet: < URL: http://www.itri.bton.ac.uk/events/senseval/>.

SUMMARY OF THE INVENTION

Preferred embodiments of the present invention include systems and methods for word sense disambiguation, including discerning one or more senses or occurrences of a term, distinguishing between senses or occurrences, and determining a meaning for a

5   sense or occurrence of a subject term. In a collection of documents containing terms and a reference collection containing at least one meaning associated with a term, the method includes forming a vector space representation of terms and documents. In preferred embodiments, the vector space is a latent semantic index vector space. In some embodiments, occurrences are clustered to discern or distinguish a sense of a term. In

10   preferred embodiments, meaning of a sense or occurrence is assigned based on either correlation with an external reference source, or proximity to a reference source that has been indexed into the space.

BRIEF DESCRIPTION OF THE FIGURES

15   Preferred embodiments of the present invention are shown by way of example and not limitation in the accompanying drawings

Figure 1 illustrates formation of the *term* by *document* matrix used in preferred embodiments of the present invention.

Figure 2 illustrates decomposition of the *term* by *document* matrix of the present

20   invention into three constituent matrices.

Figure 3 illustrates formation of the LSI matrix used in preferred embodiments of the present invention.

Figure 4 illustrates a two-dimensional representation of an LSI space according to one embodiment of the present invention.

25   Figure 5 illustrates a two-dimensional representation illustrating grouping of documents by topic within an LSI space according to one embodiment of the present invention.

Figure 6 illustrates a two-dimensional representation illustrating the location of the term *strike* in an LSI space according to one embodiment of the present invention.

Figure 7 illustrates a two-dimensional representation illustrating the distribution of documents containing *strike* within an LSI space according to an embodiment of the present invention.

Figure 8 illustrates a two-dimensional representation of an LSI space illustrating clustering of occurrences of *strike* according to one embodiment of the present invention.

Figure 9 illustrates a two-dimensional representation of an LSI space illustrating the implicit location of derived word senses according to one embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

The intended sense or meaning for a term is best determined considering the context in which it is appears. Preferred embodiments of the present invention make essential use of a vector space representation of the text passages, *i.e.* textual context, in which the terms to be disambiguated appear.

In the generation of the vector space, the following pre-processing steps may be applied to the text:

- removal of frequently-occurring words such as *the, and,* and *of,* that have little contextual discrimination value, *i.e.*, stop words;

- treating certain combinations of terms (*e.g., United States*) as a phrase;

- splitting of hyphenated terms;

- word stemming, *i.e.*, reducing words to a root form (*e.g., clears, cleared, clearing* all being reduced to the stem *clear*).

The extent to which any or all of the above pre-processing steps are applied will depend on the application.

Although other vector space representations could be used for this purpose, the technique of Latent Semantic Indexing (LSI) provides a vector space that is provably optimal in key respects. The LSI technique (including singular value decomposition and dimensionality reduction as described herein) provides a method for extracting semantic

5   information that is latent in a collection of text that is susceptible to a high degree of automation. This technique can create a full index, i.e., an LSI vector space, of a collection of documents without significant human intervention. The basic technique is described in an article by Deerwester and in U.S. Patent Number 4,839,853 (the '853 patent). Deerwester, S., et al., Indexing by Latent Semantic Analysis, Journal of the

10   American Society for Information Science, 41(6), pp. 391-407, October, 1990. The optimality of this technique is shown in Ding, C., A Similarity-based Probability Model for Latent Semantic Indexing, Proceedings of the 22nd Annual SIGIR Conference, Berkeley, CA, August, 1999. The LSI technique has been shown to be of considerable value as an approach to text retrieval.

15   The LSI technique starts with a collection of text passages, typically referred to in the literature as *documents*. The term *document* in this case may refer to paragraphs, pages, or other subdivisions of text and not necessarily to documents in the usual sense, i.e., externally defined logical subdivisions of text. For simplicity, this disclosure follows the standard convention of referring to the text passages of interest as documents. The

20   disclosure uses *term* and *word* interchangeably as elements of documents.

The use of LSI is illustrated with reference to Figure 1. As a first step, a large sparse matrix 10 (the *TxD* matrix of dimension $m \times n$, where m is equal to the number of unique terms considered and n equals the number of documents considered) formed. Each row, e.g. 12, in the *TxD* matrix 10 corresponds to a term that appears in the

25   documents of interest, and each column, e.g., 14, corresponds to a document. Each element (i, j), e.g., 16 in the matrix corresponds to the number of times that the term corresponding to row i occurs in the document corresponding to column j. For example, in Figure 1, *able* appears one time in *Doc #1* and *acid* appears two times in *Doc #2*.

- 10 -

Referring to Figure 2, the known technique of singular value decomposition (SVD) can be used to decompose the *TxD* matrix to a product of three matrices 30, 32, 34, including a matrix 32 that has non-zero values only on the diagonal. Small values on this diagonal, and their corresponding rows and columns in the other two matrices 30, 34

5   are then deleted. This truncation process is used to generate a vector space of reduced dimensionality as illustrated in Figure 3 by recombining the three truncated matrices into *(TxD)'* matrix. The relationship between the position of terms and documents in this new vector space are subject to the same properties as in the original vector space.

In order to explain the operation of the present invention, it is useful to keep in

10   mind the following properties of the vector space:

- Every document is located at the geometric average of the terms that the document contains.

- Every term is located at the geometric average of the documents that contain the term.

15   Preferred embodiments of the present invention make use of the properties of the resultant vector space defined by the *(TxD)'* matrix to determine the probable context within which terms are used in documents of interest. It should be noted that the use of the specific vector space generated using the LSI technique is not absolutely essential. In an alternate embodiment, a person skilled in the art could make use of slightly different

20   vector space representations of the documents. However, it has been demonstrated that the LSI space, as characterized by the *(TxD)'* matrix, is an optimal semantic subspace based on maximum likelihood estimation.

The present invention exploits the properties of the LSI vector space in order to determine the probable context of terms and to disambiguate them.   As   an

25   example, consider an LSI space in which a collection of English-language documents covering a broad range of topics has been indexed. Such a space typically would have several hundred dimensions. Figure 4 is an illustration of such an LSI space 110 in a

- 11 -

greatly simplified manner, *i.e.*, within the limitations of a two-dimensional figure. The documents 120 of the collection are distributed throughout the space 110, depending upon the collection of terms that they contain.

Documents 120 that treat similar subject matter cluster together in the space 110
5   due to correlations among the terms contained in the documents 120. In Figure 5, three such clusters 210, 220, 230 are shown for *strike* for illustration:

   • Documents that primarily deal with sports 210

   • Documents that primarily deal with labor issues 220

   • Documents that primarily deal with petroleum exploration 230

10   As part of the example, consider the distribution of the occurrences of *strike*. *Strike* is a particularly polysemous word. In this example, consider only three of its many meanings:

   • to swing at and miss, as in baseball,

   • to refuse to work, and

15   • to locate a new oil deposit.

      As indicated in Figure 6, *strike* has a location 310 in the LSI space 110 that is the vector sum of the locations of all of the documents that contain this word. For a polysemous word such as *strike*, this position is determined by the location of documents that contain all of the meanings of the word.

20   The distribution of *strike* among the documents 120 in the space 110 is represented in Figure 7. The occurrences of individual senses of *strike* are strongly correlated with the subject matter of the documents:

   • In the region of the LSI space 110 where documents dealing with sports 210 tend to cluster, the sense of *strike* primarily is that of swinging at a ball and
25   missing.

- In the region of the space 110 where documents dealing with labor issues 220 tend to cluster, the sense of *strike* is overwhelmingly that of a work stoppage.

- In the region of the space 110 where documents dealing with petroleum exploration 230 tend to cluster, the sense of *strike* primarily is that of striking

5   oil.

The various senses of *strike* thus are highly differentiated in the LSI space 110. This is true for most all terms with more than one topic-related sense.

Preferred embodiment of the present invention now apply standard clustering

10   techniques to discern/discriminate the one or more senses of such terms. A reference on clustering techniques is: Mathematical Classification and Clustering, Nonconvex Optimization and Its Applications, Vol. 11 by Boris Mirkin. Figure 8 shows the results of applying clustering techniques to the locations of documents that contain the word *strike*. Three regions of concentration 510 (sports), 520 (labor issues), 530 (petroleum

15   exploration) are found, as well as two isolated instances 540, 550. To a first approximation, the tightness of clustering of the groupings gives us an indication of the likelihood that the clustering reflects actual differences in word sense. The specific clustering technique that would be used to determine the tightness of the clustering would depend, in part, on the application. Most of the standard techniques for clustering

20   involve a parameter that can be measured and that reflects the distinctness of each cluster. In the present example, applying a clustering threshold results in disambiguation of three principal senses of *strike* in the documents of interest.

For each cluster that is detected, we can associate an implicit position in the LSI space corresponding to the individual sense of the term of interest that dominates that

25   cluster. This implicit position is simply the geometric average of the locations of the documents from the cluster that contain the term. The document location is used because there may be more than one occurrence of the given term in a particular document. Using the document locations keeps this operation consistent with the others that have

taken place in creating the space. It is important to maintain the proper position of these disambiguated terms with respect to all the other terms. Figure 9 shows the application of this process, arbitrarily labeling the senses of *strike* in the example as *strike*1* 610, *strike*2* 620, and *strike*3* 630. These implicit positions of the individual identified word

5   senses can be used in several important ways, as demonstrated below.

In a given application, terms for which disambiguation is desired are treated in a similar fashion to that described in the example above. In general, it is not necessary to disambiguate all of the terms that occur in the documents of interest. For example, short words such as *the, to, from*, and *as* typically are not treated, unless of special interest.

10   Similarly, terms that occur only once in a collection of documents often typically are not treated, for example, in many information retrieval applications.

From the clustering alone, we can discern/discriminate one or more senses of terms of interest. But discerning/discriminating provides no direct information regarding the actual meaning associated with each sense occurrence. That is, at this point in the

15   process we have accomplished *sense discrimination* (or discernment) but have not yet achieved *sense tagging*. In some applications, it is sufficient to know that one or more senses exist.

For example, consider the case of an information retrieval system where users provide some feedback regarding documents of interest. An initial query consisting of

20   the word *strike* would retrieve documents in all three of the clusters in the earlier example. Assuming that the user indicates an interest in a document associated with one of the clusters and/or no interest in documents associated with the other two clusters. This would be sufficient information to select (or highly rank) documents that contained the desired sense of *strike*.

25   Many applications, however, require knowledge of the specific meaning associated with each derived word sense or term occurrence. Two embodiments of the present technique allow meaning to be determined with accuracy greater than that achieved by existing methods described earlier in this disclosure. Each method can be

- 14 -

applied to either a derived sense having an implicit position, or to a single term occurrence. For ease of explanation, the methods are described using derived word sense with an implicit position.

One approach is to process a subset of terms that are located near the sense for

5   which the meaning is to be determined. In preferred embodiments of the present invention that use this approach to sense tagging, some information external to the LSI space is required. Such information could consist of a small collection of text in which word senses have been tagged with meaning, as described below. Another approach is to exploit a dictionary or thesaurus, matching the sense of a term with the closest meaning

10   or the preponderant meaning of the terms near to the sense under consideration. If the documents near the implicit position of *strike*1* contained numerous instances of words such as *pitcher, mound, ballpark, etc.*, the meaning of *strike*1* could be associated with that of a swing and a miss in baseball. Similarly, a preponderance of words such as *union, grievance, walkout, etc.*, near *strike*2* would identify this sense of the word as that

15   associated with work stoppages.

An advantage of the present invention is that the context that is considered is much larger than in typical previous techniques. There is a high probability of topically unrelated terms occurring within the small windows used in prior techniques. In the present invention, a large number of terms are taken into account. The meaning to be

20   assigned to the sense of interest is determined by the preponderance of these neighboring terms. The nature of the LSI space 110 is such that this preponderance will indicate the correct meaning an acceptable percentage of the time. The number of unrelated terms accidentally occurring close to the sense of interest is typically small in comparison to the set of preponderant terms. Thus, these "noise" terms can be effectively discriminated

25   against.

Sets of terms whose co-location indicates specific meaning readily could be derived from sources such as machine-readable dictionaries, WordNet, and tagged corpora. The inherent rich context of the present invention makes the use of such

- 15 -

resources more effective than in previous techniques. The invention inherently discriminates against noise in the form of terms that appear but are not strongly associated with the latent semantic structure, and thus with the proper senses for the terms used. Moreover, preferred embodiments of the present invention derive context

5   from whole documents and even groups of documents, rather than the smaller windows of previous techniques. The presence of noise terms in the text and the absence of desired terms in the narrow windows employed have been substantial reasons why the use of such resources in the past has yielded poor results.

Another method of identifying specific meanings is based directly on tagged

10   corpora. One or more corpora are subdivided into segments of reasonable size (*e.g.*, paragraph size) and are directly indexed into the LSI space. Preferably, each corpus and the application documents are indexed at the same time. The SVD algorithm used is not incremental. New documents can be mapped into the LSI space later in a direct fashion, but their positioning in the space is slightly inaccurate. The different senses of each term

15   from each corpora are appropriately represented so as to be indexed as different terms from the point of view of the LSI index. The implicit location of a word sense derived as described above then is compared to the actual locations of the known senses of the term in the LSI space. The derived sense is assigned the same meaning as the closest established sense (from the tagged corpora) in the space.

20   The closest established sense can be determined in several ways. Consider a sense to be disambiguated, it has a position in the space based on the documents in which it occurs. This position could be assigned the meaning corresponding to:

□ the closest instance of the same term that has a tagged sense (from the corpora);

□ the tag associated with the closest centroid of multiple instances of the same term

25   that have the same sense, *i.e.*, the implicit position of that sense of the term considering only the terms from the tagged corpora;

□ the preponderant tag of the *K*-nearest instances of tagged instances of that term (this approach to classifying objects/terms is called *K-means clustering*;

□ the tag for the closest term for which a tag is known – in case the term to be disambiguated itself does not appear in the tagged corpus;

□ the preponderant tag of the K-nearest terms that have established tags (another application of K-means clustering).

In many cases, even a single example of the sense of interest in the tagged corpus is sufficient to allow such discrimination. Even in cases of fine distinctions, only a few instances are needed. In alternate embodiments of the present invention, the user is presented with a list of most likely meanings to choose from. Through exploiting the power of the vector space indexing technique, preferred embodiments of the present invention make better use of tagged corpora than previous techniques, which typically require tens of examples of each sense.

A combination of the above two techniques also could be used. This would improve both accuracy and coverage. The coverage of the tagged corpora generally will have some limitations. However, using the tagged corpora is the most accurate approach. A good tradeoff would be to use the tagged corpora to determine the proper tag for those terms that appear in the corpora and using the dictionary, or other auxiliary resource to tag the remaining terms.

A bootstrap technique could be applied to new terms that do not occur in the supporting resources. For example, the word *bunting* might not occur in the resources used. Application of the overall technique described above might yield a clustering of the term *bunting* in two regions of the space. Following the notation used above, the clusters are labeled as *bunting*1* and *bunting*2*. As before, the senses have been discriminated, but not yet tagged with meaning. Next, the degree of overlap of the cluster regions of *bunting*1* and *bunting*2* with other cluster regions associated with terms whose sense had been tagged is calculated. The sense of each variant of *bunting* is then be assigned as comparable to the sense of the term with the greatest degree of cluster overlap. The cluster for the term *bunting*1* might, for example, most closely overlap the cluster for the term *strike*1* above. In that case, the sense of *bunting*1* is identified as

one associated with sports, or more specifically, with baseball. Continuing, the cluster associated with *bunting*2* might most closely overlap that of the sense of the word *decoration* as related to parties, thus supplying an identification of that sense. It is important to note that the precision of labeling obtained in this manner is likely to be

5   adequate for nearly all practical applications of word sense disambiguation.

The above approaches allow disambiguation of terms that occur infrequently and thus do not cluster strongly. In fact, a high-probability disambiguation can be carried out in this manner for terms that occur only once in a collection. This also applies to isolated instances of terms lying outside of the derived clusters, such as the two instances of the

10   word *strike* in documents at the bottom of Figure 9. Associating the sense of the nearest cluster in most cases will be correct.

Preferred embodiments of the present invention allow for tuning of the degree of sense separation that is achieved. There are several important parameters that can be set to control the process. Two of these are the dimensionality of the vector space and the

15   threshold for clustering.

The number of dimensions to retain in the LSI space is a parameter that is set before the space is constructed. Dimensionality of the LSI space is typically selectable in existing LSI software. In general, the larger the dimensionality, the greater the degree of separation of concepts in the space. There is, however, a tradeoff of dimensionality

20   versus generalization capability and noise immunity. One of the useful aspects of the LSI space is that terms that are similar in meaning lie close together in the space. The higher the dimensionality, the further apart similar terms will be, leaving more possibilities for overlapping with the positions of other, less-related terms.

The threshold for defining a cluster is also a parameter in most clustering

25   algorithms. In the present invention, this parameter is set so as to achieve a desired degree of separation of senses for a given term. For example, the word *strike* used in the above examples has two dozen common meanings. For a given application, it may be sufficient to differentiate only the most widely used half dozen meanings. Other

- 18 -

applications may require identifying the detailed sense of essentially every occurrence of the term.

One of the most interesting applications of the LSI technique has been in the area of cross-lingual text retrieval. This application is described in an article by Dumais and in U.S. Patent Number 5,301,109. Dumais, S., et al, Automatic Cross-linguistic Information Retrieval using Latent Semantic Indexing, in SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval, pp. 16-23. Documents in multiple languages can be represented in a single LSI space. This fact opens up a number of interesting disambiguation possibilities. In a multilingual LSI space, the techniques of the present invention would allow disambiguation of the meaning of terms in one language based on clustering of documents and sense identifications in another language. In this manner, the utility of sense definition resources could be greatly extended. In addition, individual senses of a given term in one language often are represented by different terms in another language. The locations of these terms in the LSI space could be used in labeling the senses associated with the sense clusters developed for the given term during the disambiguation process of the present invention. This would require only the availability of a machine-readable bilingual dictionary. Such resources are becoming widely available for many language pairs. The more languages in the LSI space to which this process is applied, the more accurate the mutual disambiguation will be.

It should be noted that preferred embodiments of the present invention can differentiate among multiple meanings of terms over and above linguistic variations. For example, the technique can differentiate usage of the acronym *ELF* in the context of *extremely low frequency* communications and the same acronym being used in the sense of *Eritrean Liberation Front*. This extends to multiple contexts of a single name. For example, *Eisenhower* in the context of president can be separated from the same term, identifying the same person, but used in the sense of military commander. The extent to which such distinction is necessary or desirable will vary with the application.

Yet another advantage of the present invention is the high potential for automation of the process. The present invention relies on representation of objects of interest (terms and collections of terms) in a vector space. In preferred embodiments of the invention, construction of this representation is a highly automated process. The

5   present invention requires no other information in order to carry out sense discrimination. In order to accomplish sense tagging, there is a need for a list of tags, together with some definitions that distinguish them. In the method of the present invention, such tags and definitions can be mapped into the same vector space from readily available existing resources in a highly automated manner, providing an automated ability for sense tagging

10   as well.

In a preferred computer-implemented embodiment of the present invention, programming is used to forming an m by n matrix, where each matrix element (i, j) corresponds to the number of occurrences of term i in document j. After forming the matrix, the programming performs singular value decomposition and dimensionality

15   reduction on the matrix to form a latent semantic indexed vector space. The programming then determines at least one cluster of documents within the vector space, each cluster corresponding to a subset of documents within the space that containing a subject term. The implicit position within the vector space of at least one sense of the subject term is determined as described above.

20   The nature of the present invention conveys an advantage in that it inherently requires little maintenance. Sense discrimination is maintenance-free and sense tagging requires minimal resource input. The population of the vector space is driven almost entirely by the terms that constitute the text passages of interest. Prior methods, requiring the construction of pre-defined structures involving terms or concepts, have a significant

25   limitation in this regard. Both language and the structure of human discourse change at a much greater rate than are generally realized. New terms, such as *Reaganomics*, may enter the vocabulary of discourse relatively suddenly, enjoy a period of popularity, and

- 20 -

then decline or disappear. New terminology may be generated *en masse*, such as the
plethora of e-words generated in association with the concept of e-commerce.

In addition to vocabulary, new categories for organizing information constantly
are being created, flourishing, and, often, subsequently declining. The topic of neural

5   networks, for example, is of relatively recent origin. It provides a useful conceptual
categorization for a collection of techniques and systems that have previously been
associated with a diverse set of conceptual categories. Generation of new categories
often is associated with generation of new vocabulary, in particular new senses for
existing terms. Techniques for word sense disambiguation that rely upon pre-defined

10   constructs of terms and/or categories are fundamentally unsuited for dealing with the
changes described above. Nearly all previous techniques require such constructs.

One aspect of the present invention is that it can determine the probable existence
of multiple senses for entirely new terms. This can be exploited in some applications,
such as information retrieval, without having to explicitly determine the meanings

15   associated with the new terms. The derived multiple senses for a new term may not be
entirely accurate in a linguistic sense. For many applications, however, this is not
problematical. In retrieval applications, for example, multiple senses may readily be
collapsed based on user feedback. In general, this action will be transparent to the user.
For applications requiring the tagging of word senses, the bootstrap method described

20   above can be employed for dealing with new terms in an entirely automated fashion.
This bootstrapping capability marks a major advance in the state of the art.

The present invention also derives advantage from the fact that it is completely
independent of subject matter. Preferred embodiments of the invention are driven by the
terms and text passages that are mapped into the vector space. The pre-specified

25   structures used in prior techniques generally induce an element of domain specificity.
Typically, when such techniques are applied in new domains their performance degrades
significantly. In many technical domains, for example, terms are used as terms of art
with meanings quite different from their meanings in ordinary discourse. Prior methods

- 21 -

that employ pre-defined term/concept constructs typically perform poorly when dealing with such subject matter. The present invention, being driven entirely by the text of interest, inherently is resistant to such degradation. The sense discrimination approach employed is completely independent of subject matter. The bootstrapping method for

5   assigning senses to new terms also is independent of subject matter.

While prior techniques for word sense disambiguation deal solely with standard terms used in general discourse, preferred embodiments of the present invention intrinsically deal with character strings. Thus, it can disambiguate senses for proper names, acronyms, compound terms, and even alphanumerical strings. For instance, as

10   noted above, it discriminates between contexts where the acronym *ELF* stands for *Eritrean Liberation Front* and those in which it stands for *extremely low frequency*. This feature of preferred embodiments of the invention is an advantage of the present invention in business, government, and technical applications, where acronyms and specialized terminology are ubiquitous.

15   The fact that the LSI vector space representation is based on character strings means that the present invention is language-independent. In fact, as noted above, it can be used to disambiguate terms in multiple languages simultaneously. Prior techniques generally have been strongly language-dependent.

The present invention provides a natural approach to cross-lingual

20   disambiguation. That is, terms in two or more languages can be disambiguated, and the associations among disambiguated terms in the languages can be established. This is done automatically, without requiring translation of any of the text passages from any of the languages. This unique capability is of particular importance in applying the present invention in the area of machine translation.

25   An additional advantage of the present invention is that it is controllable in terms of the degree to which disambiguation is carried out. This is true in two respects. First, the extent of separation of senses of individual terms can be controlled by specifying the number of clusters to be generated. Controlling the extent of separation is related to, but

- 22 -

not necessarily the same as controlling the clustering threshold. In standard clustering algorithms, a user typically can set either a fixed number of clusters to generate from a given data set or a threshold that will produce a number of clusters based on the characteristics of the data. In the present invention, either approach can be taken. An analytically optimal approach has yet to be identified for determining the best choices of numbers of clusters or of thresholds. In general, the chosen parameter will be varied in tests and values that yield the best performance for a particular application will be chosen. This may be done for individual applications or for classes of applications. Additionally, this can be done on a term-by-term basis.

Second, the number of terms that are disambiguated can be controlled. In some applications, it is not necessary or even desirable, to draw very fine-grained distinctions among word senses. In many applications, it is not necessary to disambiguate every term. The present invention allows control over the extent to which disambiguation is applied to any term. For example, the n-most polysemous terms or the n-most frequently occurring non-stopwords can be designated for disambiguation.

An economic advantage of the present invention is that it exploits the LSI technique. This has two major consequences. First, by being based on the LSI technique, the present invention can take advantage of the millions of dollars that have been invested in the development of LSI software. Second, LSI is being used in a growing number of information processing applications. For any collection of documents, there are no differences in the LSI space created for other reasons, such as information retrieval, and the one employed in the present disambiguation technique. Thus, in applications where the LSI technique will be used for other purposes, the present technique can be implemented without requiring additional data structures.

The technique presented here does not require the term to be disambiguated to appear many times in either the documents of interest or in any supporting material used in defining senses. In fact, the technique is capable of determining the sense of a term that appears only once in either collection.

- 23 -

CLAIMS

I claim:

1.   1.   In a vector space representing the latent semantic content of a collection of
2.   documents, a method for discerning the presence of at least one sense of a subject term,
3.   the method comprising:
4.   determining at least one cluster of documents within the vector space, each cluster
5.   corresponding to a subset of documents within the vector space containing
6.   a subject term, and
7.   determining an implicit position within the vector space of at least one sense of
8.   the subject term, the implicit position corresponding to at least one
9.   determined cluster.

1.   2.   The method in accordance with Claim 1, wherein the vector space is a latent
2.   semantic indexed vector space.

1.   3.   In a collection of documents, each document containing a plurality of terms, a
2.   method for discerning the presence of at least one sense of a subject term, the method
3.   comprising:
4.   forming an m by n matrix, where each matrix element (i, j) corresponds to the
5.   number of occurrences of term i in document j;
6.   performing singular value decomposition and dimensionality reduction on the
7.   matrix to form a latent semantic indexed vector space;
8.   determining at least one cluster of documents within the vector space, each cluster
9.   corresponding to a subset of documents within the vector space containing
10.   a subject term; and
11.   determining an implicit position within the vector space of at least one sense of
12.   the subject term, the implicit position corresponding to at least one
13.   determined cluster.

1.   4.   In a collection of n documents and a reference collection, each document
2.   containing terms, the reference collection containing at least one meaning associated with

- 24 -

3   a term, the total number of terms occurring at least once in the document collection equal

4   to at least m, a method for determining a meaning for a sense of a subject term, the

5   subject term found in at least one document and associated with at least one meaning, the

6   method comprising:

7   forming an m by n matrix, where each matrix element (i, j) corresponds to the

8   number of occurrences of term i in document j;

9   performing singular value decomposition and dimensionality reduction on the

10   matrix to form a latent semantic indexed vector space;

11   determining at least one cluster of documents within the vector space, each cluster

12   corresponding to a subset of the document collection, each member of the

13   subset having at least one occurrence of a subject term;

14   discerning an implicit position of a sense of the subject term, each implicit

15   position corresponding to at least one determined cluster;

16   discerning at least one non-subject term within the vicinity of the implicit position

17   of the sense; and

18   assigning to the sense having a discerned implicit position, the meaning,

19   associated with the term in the reference collection, that correlates best

20   with the discerned non-subject terms closest to the implicit position of the

21   sense.

1   5.   In a vector space representing the latent semantic content of a collection of

2   documents, and in a reference collection comprising at least one meaning associated with

3   a term, a method for determining a meaning for a sense of a subject term, the subject term

4   found in at least one document and associated with at least one meaning, the method

5   comprising:

6   determining at least one cluster of documents within the vector space, each cluster

7   corresponding to a subset of the document collection, each member of the

8   subset having at least one occurrence of a subject term;

9   discerning an implicit position of a sense of the subject term, each implicit

10   position corresponding to at least one determined cluster;

11   discerning at least one non-subject term within the vicinity of the implicit position

12   of the sense; and

13   assigning to the sense having a discerned implicit position, the meaning,

14   associated with the term in the reference collection, that correlates best

15   with the discerned non-subject terms closest to the implicit position of the

16   sense.

1   6.   The method in accordance with Claim 5, wherein the vector space is a latent

2   semantic indexed vector space.

1   7.   In a collection of n documents and a reference collection, each document

2   containing terms, the reference collection containing at least one meaning associated with

3   a term, the total number of terms occurring at least once in the document collection equal

4   to at least m, a method for determining a meaning for an occurrence of a subject term, the

5   subject term found in at least one document and associated with at least one meaning, the

6   method comprising:

7   forming an m by n matrix, where each matrix element (i, j) corresponds to the

8   number of occurrences of term i in document j;

9   performing singular value decomposition and dimensionality reduction on the

10   matrix to form a latent semantic indexed vector space;

11   discerning the position, within the vector space, of an occurrence of a subject

12   term; and

13   assigning to the occurrence, the meaning, associated with the subject term in the

14   reference collection, that correlates best with non-subject terms closest to

15   the implicit position.

1   8.   In a vector space representing the latent semantic content of a collection of

2   documents, and in a reference collection comprising at least one meaning associated with

3   a term, a method for determining a meaning for a occurrence of a subject term, the

- 26 -

4   subject term found in at least one document and associated with at least one meaning, the

5   method comprising:

6   determining at least one cluster of documents within the vector space, each cluster

7   corresponding to a subset of the document collection, each member of the

8   subset having at least one occurrence of a subject term;

9   discerning an implicit position of a sense of the subject term, each implicit

10   position corresponding to at least one determined cluster;

11   discerning at least one non-subject term within the vicinity of the implicit position

12   of the sense; and

13   assigning to the sense having a discerned implicit position, the meaning,

14   associated with the term in the reference collection, that correlates best

15   with the discerned non-subject terms closest to the implicit position of the

16   sense.

1   9.   The method in accordance with Claim 8, wherein the vector space is a latent

2   semantic indexed vector space.

1   10.   In a collection of n source documents and a collection of x reference documents,

2   each document containing terms, each reference document containing at least one

3   meaning associated with a term, the total number of terms occurring at least once in the

4   combination collections equal to at least m, a method for determining a meaning for a

5   sense of a subject term, the subject term found in at least one source document and

6   associated with at least one meaning, the method comprising:

7   forming an m by $[n + x]$ matrix, where each matrix element (i, j) corresponds to

8   the number of occurrences of term i in document j;

9   performing singular value decomposition and dimensionality reduction on the

10   matrix to form a latent semantic indexed vector space;

11   determining at least one cluster of documents within the vector space, each cluster

12   corresponding to a subset of the $[n + x]$ documents having at least one

13   occurrence of a subject term;

14   discerning the implicit position of at least one sense of the subject term

15   corresponding to at least one determined cluster; and

16   assigning to at least one sense corresponding to at least one discerned implicit

17   position, the meaning of the subject term closest within the vector space to

18   the implicit position of the sense.

1   11.   The method as described in Claim 10 wherein each document in the reference

2   source corresponds to one meaning.

1   12.   In a collection of n source documents and a collection of x reference documents,

2   each document containing terms, each reference document containing at least one

3   meaning associated with a term, the total number of terms occurring at least once in the

4   combination collections equal to at least m, a method for determining a meaning for an

5   occurrence of a subject term, the subject term found in at least one source document and

6   associated with at least one meaning, the method comprising:

7   forming an m by $[n + x]$ matrix, where each matrix element (i, j) corresponds to

8   the number of occurrences of term i in document j;

9   performing singular value decomposition and dimensionality reduction on the

10   matrix to form a latent semantic indexed vector space;

11   discerning the position, within the vector space, of an occurrence of a subject

12   term; and

13   assigning to the occurrence, the meaning, associated with the subject term, closest

14   to the implicit position of the sense.

1   13.   The method as described in Claim 12 wherein each document in the reference

2   source corresponds to one meaning.

1   14.   In a collection of documents, each document containing a plurality of terms, a

2   computer-implemented method for discerning the presence of at least one sense of a

3   subject term, the method comprising:

- 28 -

4   forming an m by n matrix, where each matrix element (i, j) corresponds to the

5   number of occurrences of term i in document j;

6   performing singular value decomposition and dimensionality reduction on the

7   matrix to form a latent semantic indexed vector space;

8   determining at least one cluster of documents within the vector space, each cluster

9   corresponding to a subset of documents within the vector space containing

10   a subject term; and

11   determining an implicit position within the vector space of at least one sense of

12   the subject term, the implicit position corresponding to at least one

13   determined cluster.

1   15.   In a collection of documents, each document containing a plurality of terms, a

2 computer program product for discerning the presence of at least one sense of a subject

3 term when executed on a computer system, the computer program product comprising:

4   a computer-readable medium;

5   a matrix-forming module stored on the medium that forms an m by n matrix,

6   where each matrix element (i, j) corresponds to the number of occurrences

7   of term i in document j;

8   a singular value decomposition and dimensionality reduction module stored on

9   the medium and couple to the matrix forming module that forms a latent

10   semantic indexed vector space from the matrix;

11   a clustering module stored on the medium that determines at least one cluster of

12   documents within the vector space, each cluster corresponding to a subset

13   of documents within the vector space containing a subject term; and

14   a sense position determining module stored on the medium an implicit position

15   within the vector space of at least one sense of the subject term, the

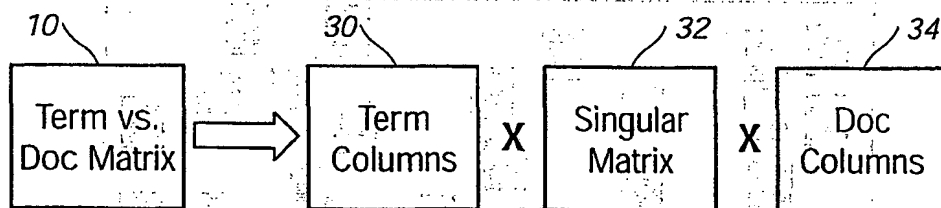16   implicit position corresponding to at least one determined cluster.

1   16.   In a collection of documents, each document containing a plurality of terms, a

2   computer program product comprising instructions that when executed perform the

3   method comprising the steps of :

4   forming an *m* by *n* matrix, where each matrix element (i, j) corresponds to the

5   number of occurrences of term *i* in document *j*;

6   performing singular value decomposition and dimensionality reduction on the

7   matrix to form a latent semantic indexed vector space;

8   determining at least one cluster of documents within the vector space, each cluster

9   corresponding to a subset of documents within the vector space containing

10   a subject term; and

11   determining an implicit position within the vector space of at least one sense of

12   the subject term, the implicit position corresponding to at least one

13   determined cluster.

	Doc #1	Doc #2	. . .	Doc #n
able	1	0	. . .	1
acid	0	2	. . .	0
. . .	. . .	. . .	. . .	. . .
<term m>	1	0	. . .	1

Term by Document Matrix

FIG. 1

| Term vs. Doc Matrix | ⟹ | Term Columns | X | Singular Matrix | X | Doc Columns |

SVD Extracts Key Relationships among Terms - Basis for Conceptual Matching
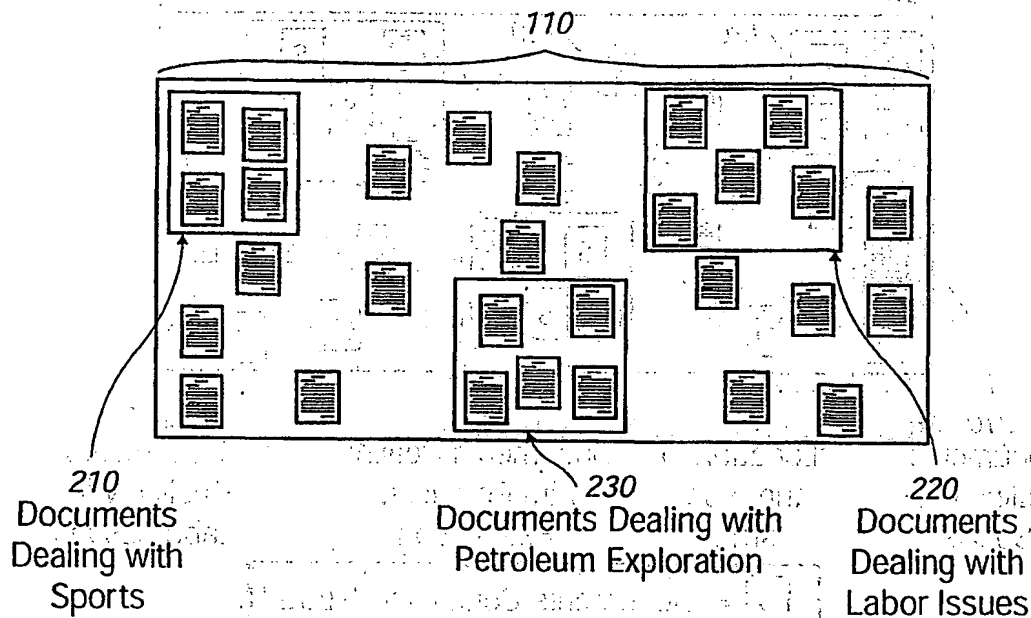
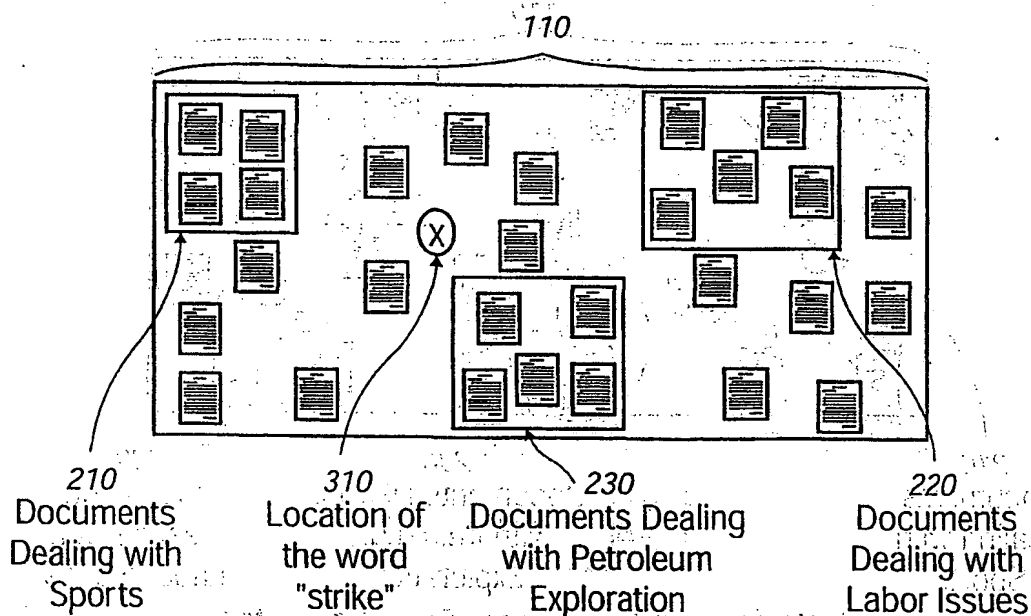Singular Value Decomposition (SVD) Operation

FIG. 2

(T x D)'
LSI Vector Space

T x D Matrix
Typically, N = Several Hundred
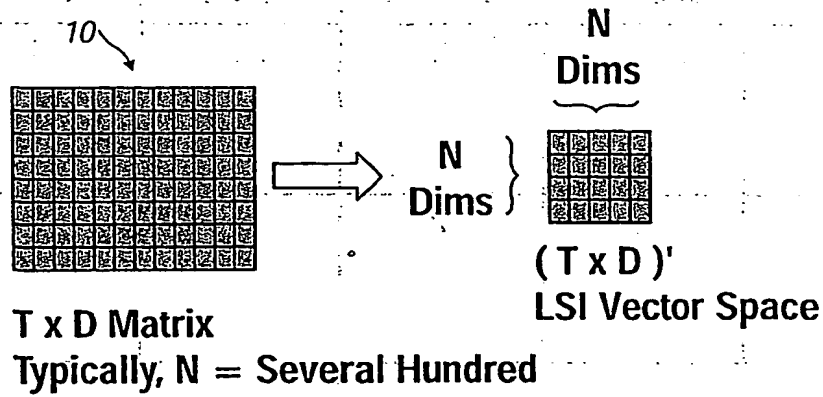
Dimensionality Reduction
FIG. 3
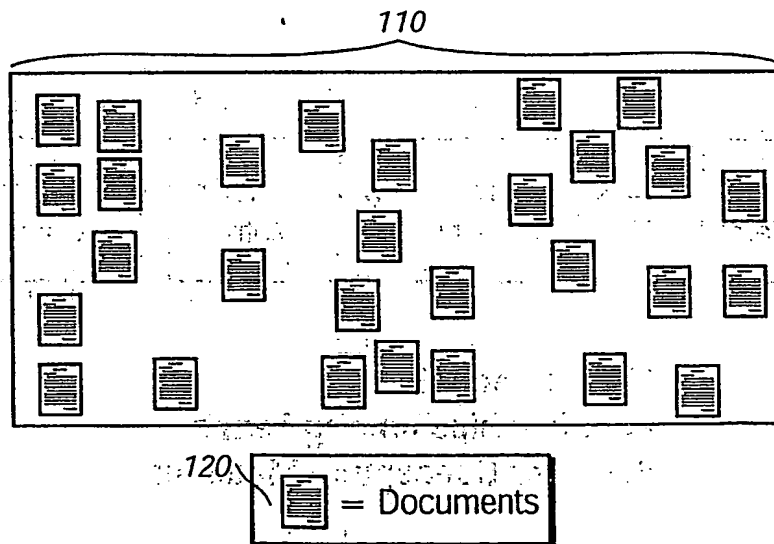
= Documents

2-D Representation of LSI Space
FIG. 4

110

210
Documents
Dealing with
Sports

230
Documents Dealing with
Petroleum Exploration

220
Documents
Dealing with
Labor Issues

Grouping of Documents by Topic in LSI Space
FIG. 5

110

210
Documents
Dealing with
Sports

310
Location of
the word
"strike"

230
Documents Dealing
with Petroleum
Exploration

220
Documents
Dealing with
Labor Issues

Location of the Word *Strike* in the LSI Space
FIG. 6
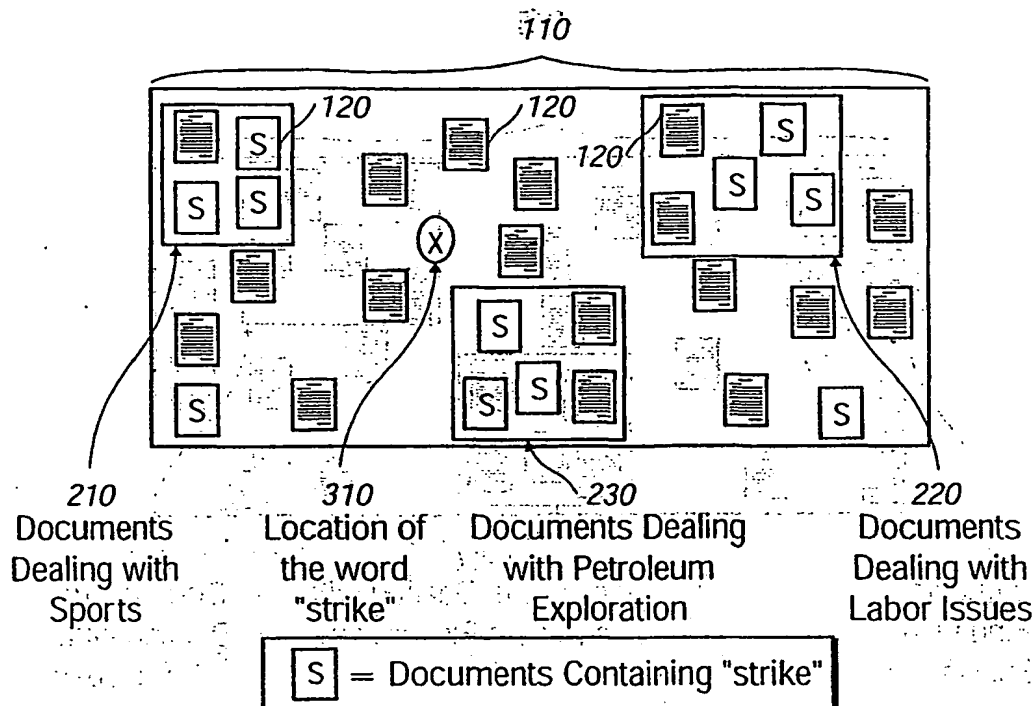
210
Documents
Dealing with
Sports

310
Location of
the word
"strike"

230
Documents Dealing
with Petroleum
Exploration

220
Documents
Dealing with
Labor Issues

S = Documents Containing "strike"

Distribution of Documents that Contain the Word *Strike*

FIG. 7

210
Documents
Dealing with
Sports

230
Documents Dealing
with Petroleum
Exploration

220
Documents
Dealing with
Labor Issues
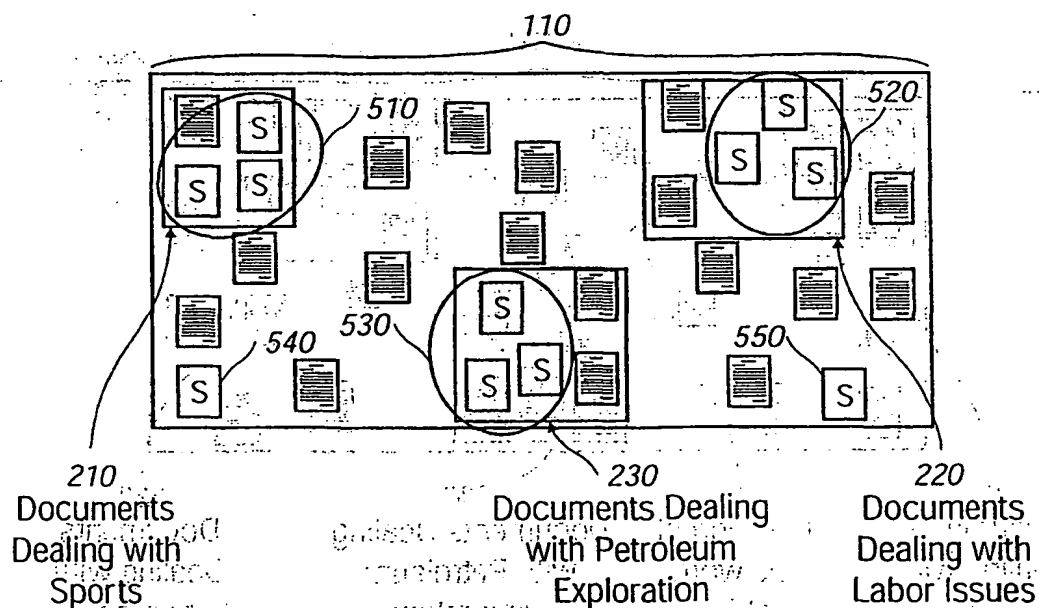
Clustering of Occurrences of the Word *Strike*

FIG. 8
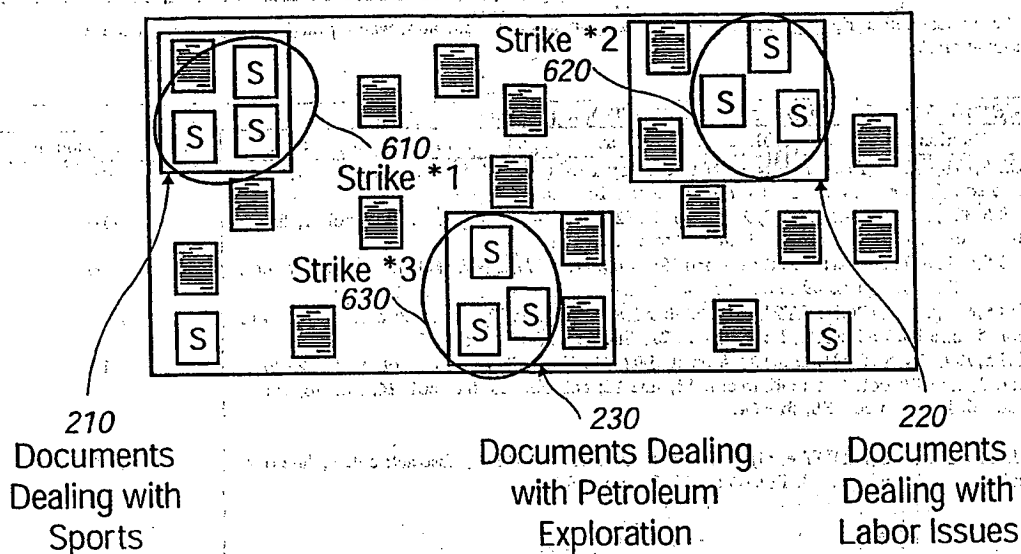
Implicit Location of Derived Word Senses
FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/26397

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7)   :   G06F 17/27
US CL   :   704/9

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 704/1, 9, 10; 707/3, 4, 5, 6, 104.1, 531, 532, 533

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,278,980 A (PEDERSEN et al) 11 January 1994 (11.01.1994), abstract; col. 1, line 8 to col. 6, line 19; and col. 15, line 54 to col. 16, line 22.	1-16
A	US 5,325,298 A (GALLANT) 28 June 1994 (28.06.1994), abstract; fig. 10; col. 1, line 10 to col. 4, line 31; and col. 4, line 67 to col. 12, line 19.	1-16
A	US 5,619,709 A (CAID et al) 08 April 1997 (08.04.1997), abstract; col. 2, line 20 to col. 3, line 42; col. 5, line 49 to col. 13, line 8; and col. 16, line 12 to col. 24, line 59.	1-16
Y	US 5,675,819 A (SCHUETZE) 07 October 1997 (07.10.1997), abstract; col. 1, line 14 to col. 5, line 15; and col. 6, line 5 to col. 21, line 13.	1-16
Y	US 5,794,178 A (CAID et al) 11 August 1998 (11.08.1998), abstract; col. 1, line 27 to col. 3, line 42; col. 4, line 42 to col. 13, line 12; col. 16, line 6 to col. 24, line 56; and col. 26, line 5 to col. 38, line 64.	1-16
A	US 5,873,056 A (LIDDY et al) 16 February 1999 (16.02.1999), abstract; col. 1, line 6 to col. 4, line 33; and col. 4, line 64 to col. 10, line 28.	1-16

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

Date of the actual completion of the international search	Date of mailing of the international search report
02 November 2001 (02.11.2001)	**16 NOV 2001**
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Joseph Thomas Telephone No. (703) 305-4760

Continuation of B. FIELDS SEARCHED Item 3:
EAST Database Searched - Files: USPAT, US-PGPUB, EPO, JPO, DERWENT, & IBM TDB

search terms: word sense, disambiguation, latent semantic indexing, matrix, vector, implicit position